On the Detection of Nontrivial and Cross Language Plagiarisms

Andreas Schmidt*† and Sören Bühler*
* Department of Computer Science and Business Information Systems,
Karlsruhe University of Applied Sciences
Karlsruhe, Germany
Email: andreas.schmidt@hs-karlsruhe.de, soeren.buehler@gmail.com
† Institute for Applied Computer Science
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: andreas.schmidt@kit.edu

***Abstract*—In this paper, we present a new approach of plagiarism detection for strongly obfuscated or even translated plagiarism, which are difficult to detect. Based on our concept, we build a prototype system and show the efficiency of our approach on different real world scenarios like wikipedia, automatically translated documents, as well as human translated books from the Gutenberg project.**

Keywords*–*Plagiarism detection; obfuscated plagiarism; cross language plagiarism; compresse bitvector; taxonomies.

I. Introduction

Plagiarism can occur on different levels. First of all, there is the one to one copy of the original text, without any further manipulation. This is also known as copy & paste plagiarism. Next, we have the syntactical obfuscated plagiarism. In this case, the words in a sentence are rearranged, or sentences are merged or splitted. On the next higher level, semantical relations between words are used. So, for example, single words can be replaced by their synonyms or hyperonyms. Another possibility is the translation of a document into another language. The last possibility is the theft of ideas. In this case, the author of a plagiarism uses foreign mental effort and claim it as its own. In this case, the concepts plagiarized were formulated in the plagiarists own words.

In our research, we focus on highly obfuscated plagiarism types. In former work [1], we developed a concept for plagiarism detection based on compressed bitvectors, word taxonomies and a graphical representation of the results using heatmaps. In the present work [2], these concepts have been concretized and extended and a prototype has been developed, which impressive shows the efficiency of our concept dealing with highly obfuscated plagiarisms.

The rest of the paper is structured as follows. In Section II the key concepts, as already described in [1] and extended in [2] are presented. Then, in Section III a number of significiant results are presented. Starting from the simpliest case of a plagiarism, the copy & paste plagiarism, different levels of obfuscation, inclusive the "theft of idea" plagiarism are presented, all based on real world data, which show the applicability of our approach.

II. Key Concepts

A. Representation of Text Fragments

Every text fragment is represented as a bitvector. The number of bits is determined by the size of the language (every bit represents a word from the language). But the size is not a critical value, because the bitvector can be compressed very well, as we have shown in [1]. One critical point is the order of the words. By sorting the words according to their frequency, we can later identify different regions of interest.

B. Size of Text Fragments

To determine a good size of our text fragments, we run a number of experiments with different text fragment sizes. To control the result, we use the PAN (**P**lagiarism Analysis, **A**uthorship Identification, and **N**ear-Duplicate Detection) document collection [3]. This collection is an evaluation framework for plagiarism detection with a known set of different plagiarisms. With this background information we run our plagiarism detection system on the PAN corpus with different sizes for the text fragments. Additionally, we use different similarity measures [4] like Jaccard, Dice, Overlap and Cosine coefficient. The best results were obtained by using a fixed fragment size of 55 words for all the used simlarity measures.

C. Similarity measure

So far, we used the pure Jaccard measure to calculate the similarity between text fragments. Because we also want to incorporate synonyms and hyperonyms in our systems, we must adapt the similarity measure accordingly. This can be done by introducing an additional bitvector A', which handles all the synonymes and hyperonyms in A. The modified Jaccard coefficient looks like in (1):

$$Jaccard'(A,B) = \frac{|A \cap B| + |A' \cap B|}{|A \cup B|}. \tag{1}$$

The synonyms and hyperonyms were determined with the help of the WordNet library [5].

D. Integration of a Weighting Factor

As described before, the ordering of the words in the bitvector results from the frequency of the words in the language. We additionally introduce a weighting, by splitting the bitvector of every text fragment into multiple parts. Thereby, we are able to give the words which occur rarely, but are therefore more relevant, a higher weight compared to more frequent words with lower relevance. Finally, we build five categories of words, with increasing weight for the words in the higher categories. The detailed formula is presented and explained in [2].

ISBN: 978-1-61208-408-4

E. Visualization

Instead of facing text fragments from the suspicious document and text fragments from the candidate set, which seem to be similar in some sense, we provide a graphical representation, abstracting from textual representation. In contrast, we are using heatmaps to express the similarity between documents. A heatmap is a two dimensional representation (a matrix), where the x and y-axis represent the text fragments from the suspicious document and a candidate document. Each field in the matrix now represents the calculated similarity value between two text fragments. The value is presented by a color gradient from white to red. A white color means, that there is no similarity between the text fragments, a red value indicates a high similarity. Figures 1 and 2 show examples of such heatmaps.

III. Experimental Results

In this section, we present a number of results. We start with the simples plagiarism, the copy & paste and continue to the more complex ones.

A. Copy & Paste Plagiarism

In this example, we use the speech of president Barrack Obama about "net neutrality" [6] as source for our plagiarism, take two fragments out of this document and inserted them into another document from the White House about the NSA scandal. The result is shown in Figure 1 on the left side. The two inserted fragments can easily be identified by so called "plagiarism lines". If you take a closer look at the heatmap, you can see, that not only the points on the well visible diagonal lines have a much higher similarity values, compared to the rest of the document, but also the values which are close to the plagiarism line. The reason for this behaviour is a partial overlapping of text from adjacent fragments. Nevertheless, the plagiarized parts can easily be identified.

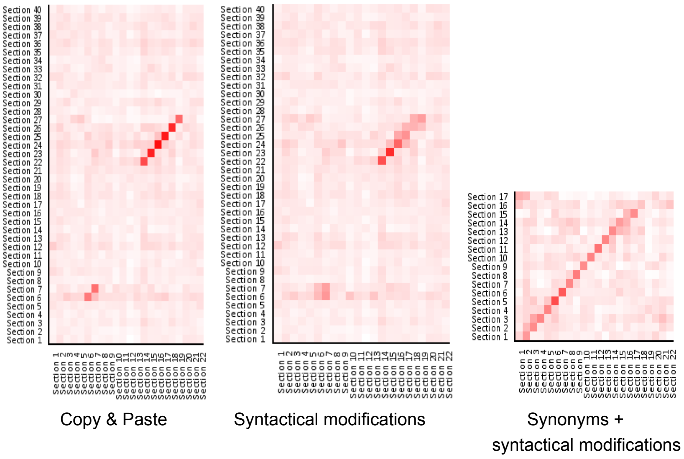

Figure 1. Results for experiments 1-3.

B. Syntactical Changes

Next, we modified the previous used copy & paste plagiarism and manipulated the plagiarized parts manually by changing the syntax (reconstruction of sentences). The result is shown in the middle of Figure 1. Compared to the simple copy & paste plagiarism the plagiarism lines are more blurred, but still very easy to identify.

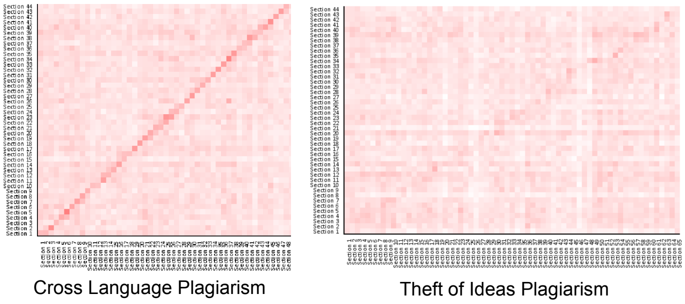

Figure 2. Heatmaps of experiments 4 and 5.

C. Use of Synonyms & Syntacical Changes

In this example, we also exchange words by their synonyms. Concretely, we again take the document about "net neutrality" and use the Google translator [7] to make a number of consecutive translations into different languages. So, in this case, we start with Obama's speech about "net neutrality" and translate it first to German, then the result to French, Spanish and in a last step back to English. Then we compare the original document and compare it with the document we achieved by a number of successive translation steps. The result can be seen on the right side of Figure 1. An inspection of the orginal and the translated text shows, that a big number of differences exist. So, for example the final text only consists of 17 text fragments compared to 22 fragments of the original text. Nevertheless, our algorithm is able to clearly detect the plagiarism.

D. Cross Language Plagiarism

In contrast to the last experiment, where an automatic transation was generated, we use a human translation in this experiment to detect cross language plagiarisms. We use books, which were available in multiple languages, from the Gutenberg Project [8]. So, in a first step we translated the German version of the book *"The adventures of Tom Sawyer"* from Marc Twain with the Google translator into English and than compared the translated version with the original English version of the book. On the left side of Figure 2 you can see the result. Also in this case, where the result of a human translator is compared with an automatic translation the plagiarism line is clearly visible.

E. Theft of Ideas

In our last example we want to test our algorithm against "Theft of Idea" plagiates. This is the most difficult plagiarism format to detect. A prerequisite was, that the two documents must be written from different authors and don't be translated one from the other. As test documents, we choose the description (first 17 episodes) of the sitcom "Big Bang Theory" from wikipedia and compare it with the correponding description found from the Internet Movie database (IMDb).

So, both authors describe the content from their perspective, with the common background that they have seen the same 17 episodes. The result of the comparision can be seen on the right side of Figure 2. Here, we can't see any more a clear line in our heatmap, but still an diagonal area with higher values.

IV. Related Work

Gottron described in [9] an approach which can be compared with our visualization technology. He also used a two dimensional representation to compare a suspicious document with another document from the candidate set. In contrast to our work, only exact matches can be visualized. Like in our approach, the appearance of diagonal lines indicate the presence of a plagiarism. The "Bag of Word" model for storing text fragments is quite common in text retrieval and plagiarism detection. This is typically done with the vectorspace model [4]. In contrast (or extension) to this approach, we further relax the information about the content of a text fragment, by only storing the information if a word appears in a fragment, but not how often it appears. This relaxation makes our approach robust against syntactical changes and allows us to use compressed bitvectors instead of integer arrays as typically used in the vectorspace approach.

V. Conclusion

We demonstrated the efficiency of our algorithm to detect obfuscated plagiarisms. The algorithm is based on the concept of storing text fragments in form of a compressed bitvector and perform similarity operations on these bitvector using an adapted version of the Jaccard coefficient. The coefficient was adapted to also support the inclusion of synonyms and hyperonyms as well as to allow a weighting of the words in the vocabularity.

Based on our tests, we identified a text fragment length of 55 as optimal. This can probably be extended by also incorporating the concept of sentences and allowing to vary the length of text fragments in a range between 40 and 60.

Another interesting research direction is the use of our approach for the "Source Retrieval" part of a plagiarism detection system. In this case, a much larger size of the text fragments should be used and also sophisticated indexes to limit the number of similarity tests must be established.

References

[1] A. Schmidt, R. Becker, D. Kimmig, R. Senger, and S. Scholz, "A concept for plagiarism detection based on compressed bitmaps," in DBKDA'14: Procceedings of the Sixth International Conference on Advances in Databases, Knowledge, and Data Applications. IARIA, 2014, pp. 30–34.

[2] S. Bühler, "Konzeption und Realisierung eines Systems zur Erkennung verschleierter Plagiate (english title: Conception and Implementation of a System to Detected Obfuscated Plagiarisms)," Bachelor's Thesis, Department of Informatics and Business Information Systems, University of Applied Sciences, Karlsruhe, Jan 2015.

[3] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An evaluation framework for plagiarism detection," in Proceedings of the 23rd International Conference on Computational Linguistics: Posters, ser. COLING '10, 2010, pp. 997–1005.

[4] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. New York, NY, USA: Cambridge University Press, 2008.

[5] "JWNL - Java WordNet Library - Dev Guide (14.12.2007)," http://jwordnet.sourceforge.net/handbook.html, [accessed 2-April-2015].

[6] B. Obama, "Speech about Net Neutrality," Nov 2014, http://www.whitehouse.gov/net-neutrality, [accessed 2-April-2015].

[7] 2015, https://translate.google.com.

[8] "Gutenberg project," https://www.gutenberg.org, [accessed 24-Jan-2015].

[9] T. Gottron, "External plagiarism detection based on standard ir technology and fast recognition of common subsequences - lab report for pan at clef 2010." in CLEF (Notebook Papers/LABs/Workshops), M. Braschler, D. Harman, and E. Pianta, Eds., 2010.